あなたの声を「すぐそば」品質で聴くAI ――遠くからでも近接マイク品質で混ざった音を聞き分ける革新的音響処理技術

話者から離れたマイク（遠方マイク）で音声を収録すると，残響や他の話者の音声，背景雑音などが混在するため，音声は聞き取りにくくなり，音声認識などの性能も劣化します．本稿では，複数の遠方マイクで収録した音から，話者の近くのマイク（近接マイク）で収録したような高品質な音声を取り出す音声強調の最新技術を紹介します．残響抑圧，音源分離，雑音抑圧を全体最適かつ高速に実現する「統一モデル」，少数マイクで高品質な処理が可能な「スイッチ機構」，さらに，深層学習に基づく音声強調（SpeakerBeamなど）との連携について述べます．

中谷 智広　池下 林太郎
加茂 直之　木下 慶介
荒木 章子　澤田 宏
NTTコミュニケーション科学基礎研究所

はじめに

近年，スマートフォンによる音声認識やヘッドセットを用いたリモート会議など，話者の口元近くにおかれたマイク（近接マイク）で収録した高品質な音声に基づく音声アプリケーションが広く利用されています．一方，今後，AI（人工知能）が，より深く私たちの生活に溶け込み，身近（＝「すぐそば」）な存在になるためには，日常生活の中で，必ずしもマイクの「すぐそば」で話されていない音声をも同等の品質で扱えるようになることが求められます．しかし，話者から離れたマイク（遠方マイク）では，壁や天井からの反射である残響，複数の話者の音声，背景雑音などが混ざってしまうため，収録音声の品質は著しく劣化し，音声アプリケーションの性能も大きく低下します．この課題を解決するために，遠方マイクで収録した音から，近接マイクで収録したかのように高品質な各話者の音声を抽出する音声強調技術の研究を進めています．特に，本稿では，単一マイクよりも高品質な処理が可能な複数マイクを用いる音声強調（複数マイク音声強調）の最新技術を紹介します．

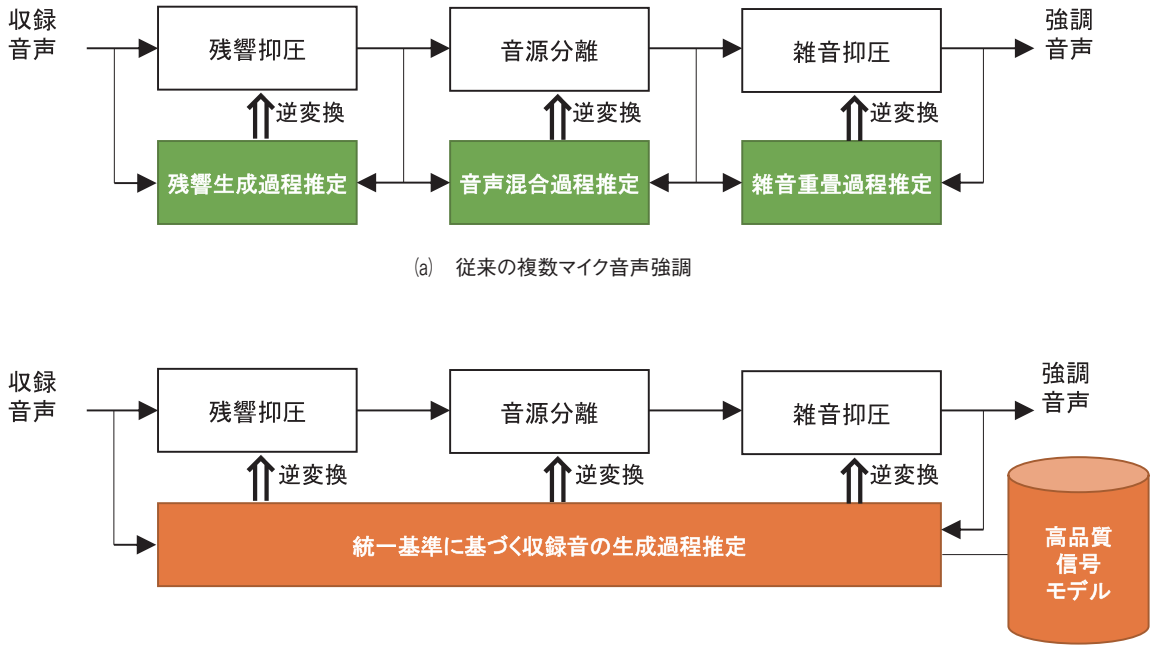
近接マイク品質の実現のための課題

収録音から近接マイク品質の音声を抽出するには，残響抑圧，音源分離，雑音抑圧の3つの処理を行うことが必要です．残響抑圧により，遠くで響いている印象のぼやけた音声を，すぐ近くにいる印象のはっきりした音声に変えます．さらに，複数の音声や背景雑音が混在している場合は，音源分離や雑音抑圧により個々の音に分解します．これら3つの処理を同時に高精度に行うことで，近接マイク品質の1人ひとりの音声に分けることができます．

従来の複数マイク音声強調では，この残響抑圧，音源分離，雑音抑圧の各課題に対し，音が音源からマイクに伝播し混合する各過程（収録音の生成過程）を推定し，その逆変換を適用することで，各処理を実現していました（**図1**(a)）．具体的には，残響が壁や天井に反射してマイクに到達する過程，複数の音声が各方向から到来し混合する過程，雑音が全方向から到来し重畳する過程をそれぞれ推定し，各逆変換を行っていました．

例えば，世界で初めてNTTが実現した残響抑圧法WPE（Weighted Prediction Error）[(1)]を用いれば，収録音中に雑音が含まれていない前提の下，どんな環境で収録されたかを知らなくても（すなわち，ブラインド処理で）収録音の残響過程を高精度に推定し，ほぼ完璧な残響抑圧を実現できていました．また，NTTをはじめ世界中で活発に研究が進められてきた独立成分分析[(2), (3)]を用いれば，収録音中に残響が含まれていない前提の下，ブラインド処理により高精度な音源分離が実現できていました．

しかし，従来技術では，この3つの処理を同時に行えないという問題がありました．つまり，雑音，残響，複数音声が混在する収録音からその生成過程のすべてを同時に推定し，その全過



(a)　従来の複数マイク音声強調

(b)　統一モデルに基づく複数マイク音声強調（提案法）

図 1　複数マイク音声強調の従来法と提案法

程の逆変換を行うことはできませんでした．その結果，各処理を順番に適用するしかなく，例えば，最初の残響抑圧では，雑音が含まれていないという前提で処理されるため，高精度な処理はできませんでした．また，音源分離や雑音抑圧では，残響が完全に抑圧されているという前提で処理されるため，最善の処理性能を得ることはできませんでした．このため，全処理を組み合わせたうえでベストの性能を引き出すこと，すなわち，全体として最適な処理を行うことはできませんでした．

私たちの日常生活環境において遠方マイクを用いて収録した音声には，必ずといっていいほど，残響，複数音源，雑音のすべての問題がつきまといます．このため，この 3 つの処理を全体最適なかたちで実現することは，音響処理における重要な未解決課題でした．

残響抑圧，雑音抑圧，音源分離の統一モデル

これに対し，私たちは，3 つの処理を全体最適なかたちで実現できる統一モデルを考案（世界初）しました[4],[5]．統一モデルでは，まず，近接マイク品質の音声や雑音が満たすべき一般的な性質を数理的にモデル化します．そして，3 つの処理を組み合わせた結果として得られる音がこの性質をもっともよく満たすようにするという「統一基準」に基づき各処理を最適化することで，全体最適な処理を実現します（図 1 (b)）．その結果，例えば，遠方マイクで収録した音声の認識性能を大幅に改善できるようになりました（**図 2** (a)～(c)）．

図 3 に，近接マイクで収録した 2 つの音声と雑音，およびそれらを遠方マイクで収録した音のスペクトログラムを示します．図から分かるように，近接マイクで収録した音声は，局所的な領域に音が集中するスパース（疎）な信号で，かつそれらが時間的に変化する非定常信号です．また，雑音は音がより広いエリアに広がったデンス（密）な定常信号です．これに対し，遠方マイクで収録した音声は，雑音，残響，複数音源が混ざったことで，近接マイクで収録した音声と比べると，よりデンスな非定常信号になるという特徴を持っています．

統一モデルでは，これらの音の特徴

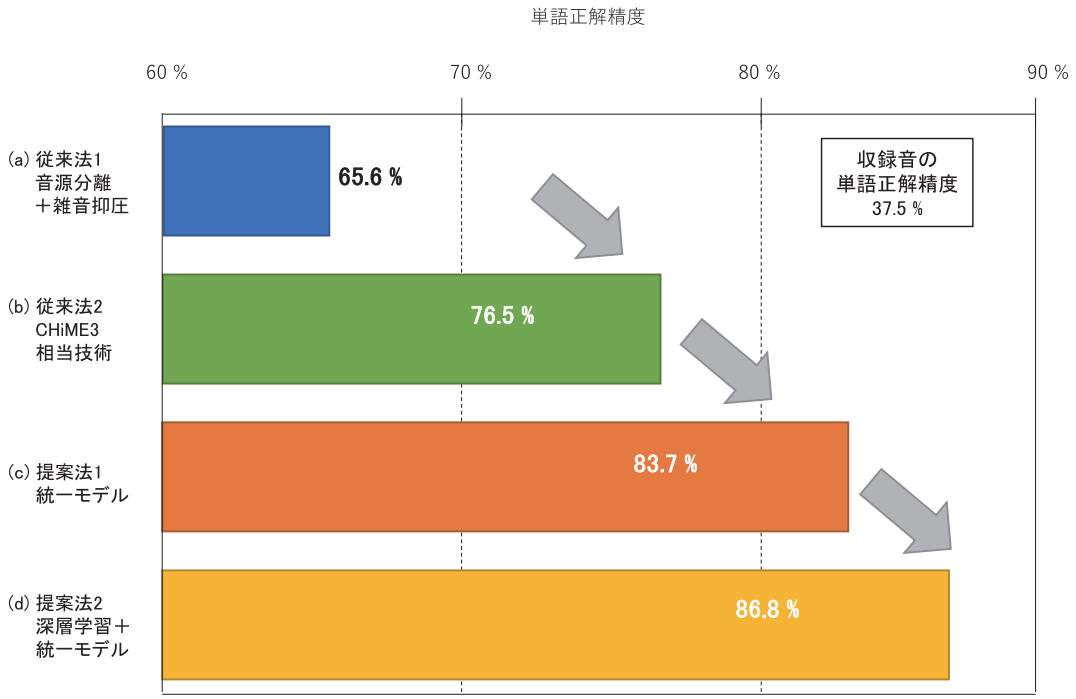


図2　複数マイク音声強調による認識率の改善

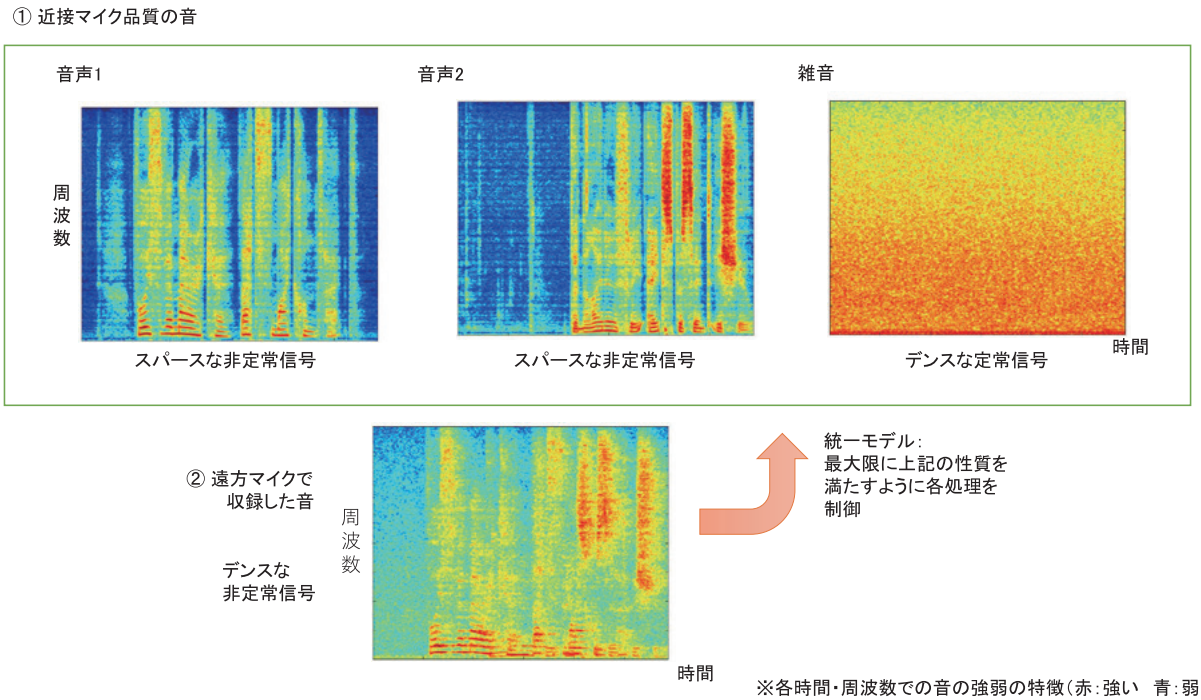
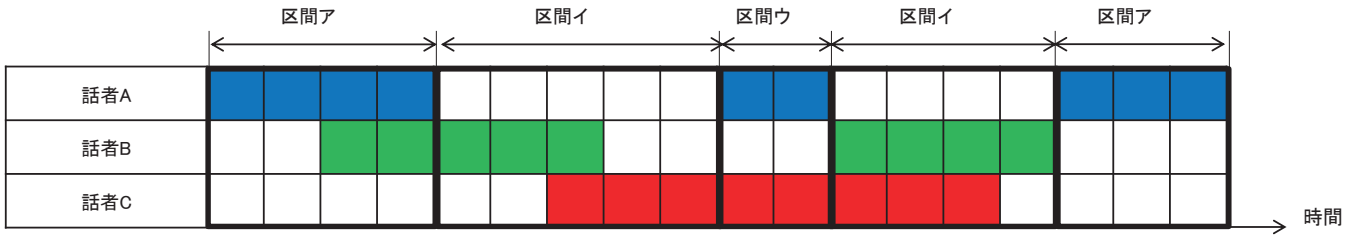


図3　近接マイクと遠方で収録した音のスペクトログラム



ア，イ，ウの各区間では、話者数は2以下になっている

図 4　3 人の会話における各者発区間の例

の違いを利用します．すなわち，残響抑圧，音源分離，雑音抑圧を適用した結果の音が，近接マイク品質の音声や雑音の特徴を最大限に満たすように各処理を制御します．例えば，残響抑圧では，音源分離や雑音抑圧と組み合わせた結果の音がもっとも近接マイク品質を満たすように，残響の生成過程とその逆変換を推定し，適用します．音源分離や雑音抑圧も，同様に，他の処理と組み合わせた結果，もっとも近接マイク品質を満たすように音の生成過程とその逆変換を推定し，適用します．その結果，近接マイク品質を実現するという目的において，すべての処理を組み合わせたうえで，全体最適な処理が行えるようになりました．

統一モデルに基づく複数マイク音声強調の処理の高速化についても，これまでに，多くの成果が得られています[6]，[7]．例えば，図 2 の実験の処理（8 本のマイクを用いて残響抑圧，音源分離，雑音抑圧の全体最適化）は，Linux計算機を用いて実時間以内で完了できるところまで高速化されています．また，残響が少ない環境で，背景雑音から 1 人の話者の音声をブラインド処理で抽出する問題に限定すれば，組込みデバイスを用いる場合でも，リアルタイム処理が可能な程度まで計算量を削減することができます．

少数マイクで高精度な推定を可能にするスイッチ機構

統一モデルを応用して，比較的少数のマイクでも高精度な推定を可能にする技術がスイッチ機構です[8]，[9]．従来，高精度な複数マイク音声強調を実現するためには，収録音に含まれる音源の数と比較して十分に多くのマイクを用いることが必要でした．これは，複数マイク音声強調を現実の問題に適用するのを難しくする課題でした．この課題を解決するために，スイッチ機構を導入することで，少数のマイクでも比較的高精度な処理が実現できるようになります．

スイッチ機構のアイデアを説明します．スイッチ機構では，収録音に多くの音源が含まれる場合でも，収録音を短い時間区間に分ければ，同時に音を発している音源の数を少なくできることに着目します．これを，**図 4** を用いて説明します．図 4 は，横軸を時間にとって，3 人の話者のそれぞれがいつ話していたかを各色付きの横バーで表しています．これに対し，図で示した短時間区間に分けると，全体では 3 人の話者がいるにもかかわらず，各時間区間は 2 人しか話していないようにすることができます．そして，話者数が少なくなった短時間区間ごとに処理を切り替えて（スイッチして）複数マイク音声強調を適用することで，少数のマイクでも比較的高精度な処理が実現できるようになります．

統一モデルに基づく音声強調にスイッチ機構を導入する場合，上記の時間区間分けを含んだ処理の全体を最適化できるという特長があります．すなわち，短時間区間に分ける処理と、区間ごとに複数マイク音声強調を行う処理に対し，それらを組み合わせた結果得られる音声が近接マイク品質を最大限に満たすように，スイッチ機構と複数マイク音声強調の両方を同時に最適化することができます．

音響処理の基礎技術としての統一モデル

前述したように，統一モデルを用いることで，これまで試行錯誤的に組み合わされてきた音声強調の 3 つの処理に対し，理論的にも実用的にも優れた

統合指針を与えることができます．また，スイッチ機構のように，より複雑な処理を組み合わせていくうえでも，統一モデルは，全体最適化を実現する仕組みを与えることができます．統一モデルは，今後，音響処理技術がさらに発展していく中で，その基礎を与える技術として広く利用されていくことが期待されます．

今後の発展：深層学習との最適な統合

複数マイク音声強調が発展していくうえで，音声強調のもう１つの重要なアプローチである深層学習との連携はとても大切です．深層学習では，SpeakerBeam[10]が実現した声の特徴に基づく選択的聴取や単一マイクによる音声強調など，複数マイク音声強調では困難な処理が実現できる一方で，残響があると処理音質が悪くなる，音声認識性能の改善は限定的であるなどの課題があります．これらの特徴は，複数マイク音声強調と相補的であり，どちらも欠くことができない技術です．例えば，複数マイク音声強調では，たくさんの音が混ざっているなかから目的話者のみを抽出することは困難ですが，深層学習の力を借りることでその課題を乗り越えることができます．一方，深層学習音声強調が推定した音声は，そのままでは音質が悪く，音声認識の性能もあまり改善できませんが，複数マイク音声強調を用いて近接マイク品質にすることで，複数マイク音声強調単体で処理する場合よりもさらに性能を向上させることができるようになります（図２(d)）．今後，両者の最適な統合方法を構築することで，より高機能で高品質な音声処理が実現されていくと考えています．

■参考文献

(1) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang：“Speech dereverberation based on variance-normalized delayed linear prediction,” IEEE Trans. on Audio, Speech, and Language Processing, Vol. 18, No. 7, pp. 1717-1731, 2010.
(2) N. Ono and S. Miyabe：“Auxiliary-function-based independent component analysis for super-Gaussian sources,” Proc. of LVA/ICA 2010, pp. 165-172, Springer, St. Malo, France, Sept. 2010.
(3) H. Sawada, S. Araki, and S. Makino：“Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” IEEE Trans. Audio, Speech, and Language Processing, Vol. 19, No. 3, pp. 516-527, 2011.
(4) T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach：“Jointly optimal denoising, dereverberation, and source separation,” IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 28, pp. 2267-2282, 2020.
(5) R. Ikeshita and T. Nakatani：“Independent vector extraction for fast joint blind source separation and dereverberation,” IEEE Signal Processing Letters, Vol. 28, pp. 972-976, 2021.
(6) R. Ikeshita, T. Nakatani, and S. Araki：“Block coordinate descent algorithms for auxiliary-function-based independent vector extraction,” IEEE Trans. Signal Processing, Vol. 69, pp. 3252-3267, 2021.
(7) T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino：“Low latency online source separation and noise reduction based on joint optimization with dereverberation,” Proc. of EUSIPCO 2021, pp. 1000-1004, Dublin, Ireland, Aug. 2021.
(8) R Ikeshita, N. Kamo, and T. Nakatani：“Blind signal dereverberation based on mixture of weighted prediction error models,” IEEE Signal Processing Letters, Vol. 28, pp. 399-403, 2021.
(9) T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki：“Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms,” IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 30, pp.1032-1047, 2022.
(10) Delcroix・Zmolikova・木下・荒木・小川・中谷：“SpeakerBeam: 聞きたい人の声に耳を傾けるコンピュータ――深層学習に基づく音声の選択的聴取,” NTT技術ジャーナル，Vol. 30, No. 9, pp. 12-15, 2018.

（上段左から）中谷 智広／池下 林太郎／加茂 直之
（下段左から）木下 慶介／荒木 章子／澤田 宏

私たちは，ロボットやコンピュータなどが人間と同様に私たちの会話を理解できるようになるための研究を進めています．複数マイク音声強調は，その重要な要素技術です．信号処理と深層学習を連携・発展させていくことで，未来のAIの耳を創造していきたい．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部
信号処理研究グループ
TEL　0774-93-5020
E-mail　cs-liaison-ml＠hco.ntt.co.jp